영한 번역의 언어학적 평가 모델 연구*

- 기계번역을 중심으로 -

A Linguistic Evaluation of English-to-Korean Translation

- Centered on Machine Translation -

김 덕 봉** 조 병 은*** 김 명 철**** 권 용 현*****
(Deok-Bong Kim) (Byung-Eun Cho) (Myung-Chul Kim) (Yong-Hyun Kwon)

요 약 기계번역 품질 평가는 중대한 문제이다. 기계번역의 품질이 사용자 요구와 거리가 상당히 있는 현재의 상황에서, 기계번역 시스템의 객관적 평가는 기계번역 소프트웨어 사용자와 판매자 간의 신뢰를 구축하고 개발자들 간에 생산적인 경쟁관계를 조성하게 하여, 결과적으로 기계번역 품질의 고급화를 지속적으로 유도하는 역할을 할 것이다. 이를 위해서는 특히 언어학적 측면과 자료처리 측면에서 개선이 계속되고 있는지를 확인할 수 있도록 기계번역 시스템의 품질을 평가할 수 있는 연구가 있어야 한다. 본 논문에서는 이런 점들을 고려해 넣은 영한 기계번역의 언어학적 평가 방법을 제시하고, 이를 몇 개의 상용 기계번역 시스템을 대상으로 실험하여 실험결과를 보고한다. 이 방법은 기본적으로 언어현상과 학습수준으로 분류된 3,373 영어 문장으로 구성된 평가자료에 기반하고 있다.

주제어 기계번역, 평가자료, 언어학적 평가, 영한 번역 품질 평가

Abstract Machine translation (MT) quality assessment is an outstanding problem. In the present situation in which the quality of machine-translated products are far from the user's satisfaction, objective evaluation of MT system is a prerequisite to building mutual trust between the users and the vendors, stimulating constructive competition among the developers, and finally leading to improve the quality of MT systems. Especially there emerges a need for an intensive study on how to evaluate the quality of MT systems from both linguistic and data processing aspects, and to secure a steady improvement of the translation quality. With due regard to such points, we in this paper present a linguistic evaluation of English-to-Korean machine translation based on a test suite composed of 3,373 sentences that were classified into their linguistic phenomena and complexity levels, and report the experimental results made from several commercial MT systems.

Keywords machine translation (MT), test suite, linguistic evaluation, English-to-Korean translation quality assessment

* 이 연구는 학술진흥재단의 학제간 공동연구 지원을 받아 수행되었음.

** 성공회대학교 컴퓨터정보공학부
(152-716)서울특별시 구로구 항동 1-1
전화: 02-2610-4274
FAX: 02-2610-4390
E-mail: dbkim@mail.skhu.ac.kr
*** 성공회대학교 영어학과
전화: 02-2610-4253
FAX: 02-2610-4298
E-mail: becho@mail.skhu.ac.kr
**** 성공회대학교 컴퓨터정보공학부
전화: 02-2610-4272
FAX: 02-2610-4390
E-mail: mckim@mail.skhu.ac.kr
***** 성공회대학교 영어학과
전화: 02-2610-4255
FAX: 02-2610-4298
E-mail: yhkwon@mail.skhu.ac.kr

1. 서론

영어 중심의 인터넷 정보 사용이 보편화되고 있는 상황에서 다수의 영한 기계번역 소프트웨어들(앙꼬르, 트래니, 인가이드, Etran2000, ClickQ, EZReader, 미래번역기 등)이 시장에 나와 일반 사용자의 지대한 관심을 받고 있다. 그러나 영한 기계번역 소프트웨어 광고에서의 번역율과 사용자가 사용하면서 느끼는 실제 번역율 간에는 아직도 큰 차이가 있다. 단적으로 살펴보면 이 연구의 시발점이기도 한 1997년에 시장에 유통되던 영한 기계번역 시스템으로는 앙꼬르, 트래니, 워드체인지가 있었는데, 이들 시스템들은 당시 신문 광고에 거의 90% 이상의 번역율을 보인다고 자랑했었다. 문제는 그러한 평가들이 어떤 근거로 어떻게 나왔는지 알 수 없기 때문에 한 번역 시스템이 다른 번역 시스템에 비해 얼마나 좋고 나쁜지를 공정하게 비교하기가 어렵고, 게다가 보통 사람의 상식을 가지고 그러한 평가를 그대로 받아들인다면 우리가 이제부터는 더 이상 외국어를 배우지 않아도 될 것 같은 착각도 할 수 있었다는 점이다. 이를 반영하듯 이들 시스템들은 초기 시장 선점에도 불구하고 현재 시장에서 사라졌거나 점유 비율이 미미하다. 이와 같이 사용자와 판매자가 느끼는 기계번역의 품질에 대한 인식의 차를 줄이고 서로 신뢰를 구축하기 위해서는 결과적으로 영한 번역 소프트웨어들 간의 공정한 비교평가를 위한 연구가 필요하다.

영한 기계번역 평가에 관한 기존의 연구는 시작 단계에 있다. 기존의 연구를 평가 방법과 평가문 구성 관점에서 특징과 문제점 중심으로 살펴보면 다음과 같다. 우선, [1]에서 행한 연구를 살펴보면, 평가문은 특정 전문영역 텍스트(컴퓨터 잡지, 매뉴얼 등)의 문장들을 문장 길이별로 구성하고, 평가는 '우수(best)', '양호(good)', '부족(poor)', '실패(fail)'와 같은 4가지 등급의 번역 정확성 평가 기준과 인간 번역문에 바탕을 두어 평가문의 기계번역 결과에 대한 정확성을 평가했다. 이 연구의 특징은 프로젝트로 개발한 영한 기계번역 시스템의 성능(연구개발 성과)을 측정하기 위한 목적으로 사용되었다는 점이다. 또 평가문은 특정 영역에서 만들었기 때문에 상대적으로 일반 적용이 어려운 문제를 갖고 있다.

[2]에서는 평가문이 문장의 형식이나 문법적인 구성의 난이도, 문장의 길이 등을 대략 고려하여 구성하고, 평가는 '가장 매끄러운 번역(A)', '약간 어색한 번역문(B)', '전혀 엉뚱한 번역(C)', '영문 그대로 옮긴 문장(D)'과 같은 4가지 등급의 기준에 따라 번역의 정확성을 평가했다. 이 연구는 일반 문장을 대상으로 문장의 형식이나 문법적 구성의 난이도에 따라 구성된 200문 규모의 평가문에 바탕을 두어 여러 영한 기계번역 시스템들(워드체인지, 트래니, 앙꼬르)을 비교하기 때문에 일반성을 가질 수 있다는 점이 특징이다. 그러나 평가문의 완전성(completeness)[1]이나 체계성(systematization)[2]은 상당히 부족한 편이다.

[3]에서 평가문은 다양한 영어 문법 구조와 현상들을 기반으로 구성한 650문 규모의 평가자료(Test Suite)에서 선정하고, 평가는 6등급을 기반으로 하면서도 각 평가항목마다 가중치를 부여하고 있는 절대적인 기준(명제내용(40%), 문법구조(20%), 단어인식(10%), 완성도(20%))과 상대적인 기준(번역 속도(10%))에 따라 수행하고, 또한 번역오류의 유형과 분석을 수행했다. 이 연구의 특징은 국내에서 처음으로 평가자료를 구성하여 영한 기계번역의 유형학적 평가[3] [4][5]를 시도하였다는 점이다. 그러나 평가자료의 구성이 영한 변환 현상은 특별히 고려하지 않은 채 영어의 일반적 문법현상에만 바탕을 두고 있어 단순하고, 또 문법현상을 세분하여 구분하거나 평가문의 난이도를 구분하지 않아 체계적이지 못한 편이고, 평가기준도 복잡하여 대규모 평가 시에 비용이 많이 들 수 있는 문제를 가지고 있다.

[6]에서는 인간의 언어능력 평가와 같은 방법으로 기계번역기의 능력 평가를 시도하였다. 평가문은 기계번역기의 어휘적 능력, 구문론적 능력, 의미.화용론적 능력을 다양하게 측정하기 위하여 어휘, 관용어/숙어, 논항, 생성, 용법, 복잡문, 응용, 전처리 등 8개의 대범주와 문장 난이도(상/중/하)를 분류하여 구축한 1501문이 사용되었다. 그리고 평가는 3단계에 걸쳐 모범 번역문 없이 평가자(국어학 분야의 전문가)의 주

1) 평가문의 완전성은 평가문이 측정하려는 영역의 모든 중요 언어현상을 빠짐없이 대표하고 있어야 함을 의미한다.

2) 평가문의 체계성은 다양한 평가를 지원하기 위하여 평가문이 언어현상별, 수준별 등으로 잘 분류되어 있어야 함을 의미한다.

3) 유형학적 평가는 평가자료를 사용하여 시스템의 언어학적 적용범위를 조사하고, 분석하기 위한 평가방법으로 오류의 수보다는 오류의 종류에 관심을 더 많이 둔 평가방법이라 할 수 있다. 여기에서 평가는 평가문의 번역 결과에 대해 좋음(good), 나쁨(bad), 혹은 실패(zero) 판정에 의해 이루어진다. 이 방법의 장점으로는 시스템 개발자들이 어떤 구문들이 번역이 잘 안되고 있고, 개선이 필요한지 쉽게 파악할 수 있고, 개선 작업에 대한 효과를 쉽게 측정할 수 있으며, 시스템들간의 직접 비교가 가능하다는 점을 들 수 있다. 그러나, 평가자료를 구축할 때 언어현상별로 실제 텍스트의 빈도 가중치를 주는 것이 어렵기 때문에 평가결과가 실제 번역율과 차이가 날 수 있다.

1501문이 사용되었다. 그리고 평가는 3단계에 걸쳐 모범 번역문 없이 평가자(국어학 분야의 전문가)의 주관적 판단에 따라 문장별로 이루어졌다. 즉, 1단계 과정에서는 평가자가 해당 평가범주에 대한 만족도를 중심으로 네 등급(0~3점) 중 하나를 선택한 다음, 2단계 과정에서는 문장 전체 번역의 완성도를 고려하여 등급을 조정(1~2점까지 감점)하고, 3단계에서는 다른 평가 대상 기계번역기와의 형평성을 고려하여 점수를 조정하는 과정을 거쳤다. 이 연구의 특징은 인간과 같이 기계번역기의 질적 평가를 시도하여 직관적으로 단순성을 주고 있지만, 기계와 인간의 차이를 지나치게 고려하지 않아 다른 것들과 형평성이 맞지 않는 평가 항목들(예: 어휘, 관용어/숙어)이 들어있고, 평가 기준과 절차가 상당히 복잡하여 평가자의 주관적 측정을 객관화하는데 많은 비용이 들어갈 수 있는 문제점을 갖고 있다.

위에서 살펴본 바와 같이 영한 기계번역의 객관적 평가를 위한 연구는 아직 미흡하다. 특히 현재의 영한 번역 소프트웨어 기술은 사용자의 기대를 충분히 만족시키기에는 아직 부족한 점이 많기 때문에 개발자가 번역 품질의 고급화를 지속적으로 추구할 수 있는 평가 방법, 즉 언어학적 측면과 자료처리 측면에서의 개선을 정량적으로 평가할 수 있는 방법에 대한 연구가 우선적으로 요구된다. 이러한 요구에 따라 본 논문은 영한 기계번역 결과를 언어학적으로 평가하기 위한 방안을 제시한다. 2장에서는 영한 번역의 언어학적 평가를 위한 기반이라 할 수 있는 평가자료에 대하여, 다양한 영한 번역 평가를 위한 평가문 선정, 평가문의 언어현상과 수준 분류, 평가문의 전문가 번역문 작성 등 평가자료의 설계와 구축내용을 다룬다. 그리고 3장에서는 번역의 정확성과 이해도를 평가하기 위한 기준을 기술하고, 4장에서는 이에 바탕을 두어 실제 영한 기계번역의 언어학적 평가를 실험적으로 수행하고 이 실험 결과를 분석한다. 끝으로 5장에서는 본 연구의 결과를 요약하고, 향후 바람직한 연구 방향과 과제를 다룬다.

2. 영한 번역 평가자료의 설계와 구축

본 연구에서 영한 번역의 평가는 다음 (1)과 같이 구축한 평가자료에 바탕을 둔 유형학적 평가 전략을 사용하여 이루어졌다.

(1)영한 번역 평가자료

a. 구성 항목:

- 일련번호
- 평가문의 언어현상: 영어 고유 현상, 영한변환 특수 현상
- 평가문의 수준: 초급, 중급, 고급
- 평가문(영어문장): 3,373문
- 평가문의 모범번역문(한국어문장)
- (평가 부분 지시: 영한 변환 특수 현상 평가문에 국한)

b. 구축 규모:

- 영어 이해능력 평가문 3,073문: 초급(1,810문), 중급(1,022문), 고급(241문)
- 영한 변환능력 평가문 300문

c. 구축 예제:

0001 A01 초급
Abrams works. 아브람스는 일한다.

0002 A01 초급
She lent some money to her friend. 그녀는 친구에게 돈을 좀 빌려 주었다.

0003 A01 중급
He asked me if I had found it. 그는 내가 그것을 찾았는지 물었다.

0004 A04 고급
If these things had not been done, it is doubtful that the world would have survived this long. 만약 이것들이 이루어지지 않았었더라면, 세상이 이렇게 오래 존재할 수 있었을지 의심스럽다.

.....

3074 101 초급
The area of this room is 150 square feet. 이 방의 면적은 150평방 피트이다. (area: 면적)

3075 101 초급
His research embraced the whole area of education. 그의 연구는 교육 전 분야를 포괄했다. (area: 분야)

3076 101 초급
The area is encircled by police. 그 지역은 경찰에 의해 포위되었다. (area: 지역)

.....

본 평가자료를 설계할 때 목표로 했던 점은 다음과 같다. 첫째, 영어 평가문은 영어의 모든 문법현상을 대표할 수 있고, 문법적으로 정확하고, 수준별로 체계화되어 있는 예문들로 구성하고자 했다. 그 이유는 평가문이란 "자연언어처리시스템의 적용범위를 평가할 수 있는 기준이 되는 자연언어 문장의 모음"[7]으로 교육 관점에서 보면 시험문제와 같기 때문이다. 따라서 평가문은 시험문제라면 흔히 갖추어야 될 여러 가지 요소, 특히 완전성과 정확성과 체계성을 지닐 수 있는 일종의 표본문장의 모음이어야 한다고 보았다.

둘째, 번역은 원어와 목표어라는 두 개의 언어를 대상으로 하는, 두 언어에 대한 고른 이해를 기초로 하

는 활동이므로 평가자료는 원어에 대한 이해능력과 목표어로의 변환능력을 동시에 측정할 수 있어야 한다는 점을 반영하고자 했다. 즉, 원어인 영어의 모든 언어학적 현상을 포함하고, 목표어인 한국어로의 변환능력을 적절히 평가할 수 있는 영어와 한국어 두 언어간의 문법, 구문, 어휘 및 어법의 차이를 고려해 만든 평가자료를 구축하여 영어 번역의 언어학적 평가를 하고자 했다.

끝으로, 평가 측정의 객관성을 확보하고자 했다. 일반적으로 평가를 하기 위해서는 시험문제의 모범답안과 같이 판단의 기준이 될 수 있는 객관적 잣대가 필요한데, 이러한 잣대는 가능한 한 언제 누가 측정하더라도 문제 답안에 대하여 일정한 판정을 내릴 수 있어야 한다. 번역 평가에서는 본 연구에서와 같이 각 평가문에 대하여 전문가 번역문을 마련해 두고, 보통 이를 평가자가 번역의 정확성을 측정할 때의 판단기준으로 사용할 수 있도록 하고 있다. 이것은 평가자에게 평가 대상문에 대한 모범적인 번역 결과를 미리 줌으로써 평가자가 원어를 이해해서 판정해야 되는 부담을 줄이고, 평가자들 간의 판정 오차를 줄이기 위한 노력이라 할 수 있다.

2.1 평가문의 언어현상 분류체계

영한 번역 평가를 위한 평가문은 영어 이해능력 측정문과 영한 변환능력 측정문으로 나누어져 있다.

2.1.1 영어 이해능력 측정문

영어 이해능력 평가를 위한 문장 수집은 거의 모든 영어 문법현상들을 수용한 것으로 평가받고 있는 HP-NL 평가자료[8]의 분류에 기반하였다. HP-NL 평가자료는 1987년 6월 30일자로 공개된 것으로 Dan Flickinger, Marilyn Friedman, Mark Gawron, John Nerbonne, Carl Pollard, Geoffrey Pullum, Ivan Sag, Tom Wasow 등 8인의 언어학자들이 Hewlett Packard사에서 자연언어 처리를 위해 작성한 평가자료(Test Suite)이다. 이 평가자료는 (2)와 같이 영어의 언어학적 제 현상을 20개의 큰 항목과 77개의 세부항목으로 분류, 이를 문법적으로 오류가 없는 912개의 문장과 313개의 비문법 구문으로 구분하였고, 평가문은 문장 단위로 구성하였으며 각 항목별로 그것이 언어학적 기준으로 핵심사항(core)인지 주변적인 것(periphery)인지를 명시하여 영어의 제 현상을 모국어 화자의 입장에서 분석하였다. 이 평가문은 문장의 통사론적이나 의미론적 분석에만 그친 것이 아닌 담화에 대한 분석과 추론, 궁극적으로는 그 이상의 것까지도 평가하기 위한 의도로 제작되었으며, 따라서 광범위한 통사적, 의미론적 현상과 담화론적 현상까지를 포함하는 상당히 포괄적인 문장들로 구성되어 있다. 평가문의 스타일은 비공식적이고 타이프로 친 의사소통문에 혼하면서도 방언이나 구어체는 피한 스타일을 선택하고, 또 다른 스타일상의 특색으로는 사람과 NLP 시스템간의 상호작용도 일종의 담화로 간주, 문장과 문장 간 의존도를 포함시킨 것을 들 수 있다.

2.1.2 영한 변환능력 측정문

영한 변환능력 평가부분은 영어 어휘와 구조가 한국어로 변환하는 과정에서 얼마만큼 한국어 어법에 맞게 바뀌느냐 여부에 초점을 두어 4부분으로 분류하였다.

1. **어휘변환**: 다의어(homographs & polysemy)와 다중번역동의어(multiple translation equivalents)[4]로 문맥에 따라 다른 표현의 한국어 단어로 변환될 수 있는 영어 단어(명사, 동사, 형용사, 부사)를 대상으로 한다.

101. 명사:
예) 1. The bank opens at 10:00 am.
은행은 오전 10시에 문을 연다.
2. She stood on the near bank of the river.
그녀는 가까운 강둑에 서있었다.

102. 동사:
예) 1. He was ordered back to Japan.
그는 일본으로 돌아가라고 명령받았다.
2. The organisms are ordered according to species.
조직체는 종에 따라 배열되었다.
3. I ordered dinner for fifty people.
나는 50명분 저녁을 주문했다.

103. 형용사:
예) 1. We were free from interference.
우리는 간섭으로부터 자유로웠다.
2. The exhibition is free of charge.
그 전시회는 무료 입장이다.
3. Are you free next weekend?
다음 주말에 한가합니까?

4) 영어 단어의 의미는 모호하지 않지만 여러 한국어 단어나 표현으로 번역될 수 있는 것을 말한다. 예를 들어, 영어 단어 'wear'는 문맥에 따라 한국어로 '입다, 신다, 끼다, 차다, 걸다, 바르다' 등으로 번역될 수 있다.

(2) 영어 이해능력 평가항목 분류

A. 어휘의존도
A01. 동사의 종류
A02. 전치사 선택
A03. 격 부여
A04. 허사(유도부사)
A05. 수/관사 일치
A06. 조동사의 사용
A07. 절 보어
A08. 부분사
A09. 통제
A10. 결과문
A11. 도구문
A12. 수동구문
A13. 가성 수동 구문
A14. 계사 사용 구문
B. 평서문위치이동
B01. 도치구문
C. 관계절
C01. 제한적 관계절
D. 의문문
D01. 독립의문문
D02. 종속의문문
E. 비교상관구문
E01. 비교구문
E02. 동격구문
E03. 최상급 구문
E04. 비교급의 삭제변형
E05. 비교급의 준삭제
E06. 비교급 생략유형1
E07. 비교급 생략유형2
E08. 비교급 생략유형3
E09. 비교급 생략유형4
E10. 선도규약에 의한 제거
E11. 전치사 비교구문
E12. 비교절의 도치
E13. 비교문에서의 공백화
E14. 비교구문의 후치
F. 등위구문
F01. 이중접속 대 다중접속
F02. 'but'에 의한 접속
F03. 접속간 의존성
F04. 유사범주와 이질범주
F05. 경계교차
F06. 명사구접속과 수 일치
F07. 비성분 등위 접속
G. 조응사
G01. 자유 대용사
G02. 결속 대용사
G03. 동사구 생략
G04. 간접 의문의 단축형
G05. 의사 공백화
G06. 좌측 주변부 생략
G07. 영 보어 대용사
H. 부사수식
H01. 문장 수식 부사
H02. 태도 부사
H03. 보조부사
I. 명사구
I01. 명사 보문
I02. 제한적 수식
I03. 관계절
I04. 명사와 명사의 복합
I05. 형용사구 사용
I06. 소유격
I07. 부분사와 명사구
I08. 인명 및 직위 표현
I09. 약어 및 두문자어
I10. 날짜
I11. 지명
I12. 수
I13. 양/수의 일치
I14. 한정사
J. 외치구문
J01. 'It' 사용 외치
J02. 명사구 외치
K. 명령구문
K01. 명령문
L. 목적어 공백 보문
L01. 목적어 공백 보문
M. 분열문
M01. 'It' 사용 분열문
M02. 'Wh' 분열 구문
N. 수량 표시 구문
N01. 수량의 표현
O. 조건문
O01. 조건절의 사용
P. 부정문
P01. 부정문
Q. 시제와 상
Q01. 시제
Q02. 상
R. 서술 부가어
R01. 서술 부가어
S. 삽입문
S01. 삽입구문
T. 직접인용문
T01. 직접인용문

104. 부사:
예) 1. He has not come yet.
그는 아직 오지 않았다.
2. We may win yet.
우리는 언젠가 이길 것이다.

2. 격 변환: 단어 위치나 전치사에 의해 표현되는 영어의 격을 적절한 한국어 격조사로 변환하는 것을 평가하기 위한 것이다. 문장 내에서 다른 단어와의 관계를 나타내는 격에는 일반적으로 주격(Nominative), 목적격(Objective), 여격(Dative), 처격(Locative), 조격(Instrumental) 등이 있다.

201. 주격:
예) 1. It is snowing. 눈이 옵니다.
2. Did you telephone me yesterday?
네가 어제 나에게 전화를 했니?

202. 목적격:
예) 1. John loves Mary. 존은 메리를 사랑한다.
2. John married Mary.
존은 메리와 결혼했다.
3. Do you know who I am?
제가 누구인지 아십니까?

203. 여격:
예) 1. John sold a house to Fred.
존은 프레드에게 집을 팔았습니다.
2. Did you send the letter to your hometown?
고향 집에 편지를 보냈습니까?

204. 처격:
예) 1. John bought a house from Fred.
존은 프레드한테서 집을 샀습니다.
2. Where do you work? 어디서 일하십니까?

205. 조격:
예) 1. I came to Korea by an airplane.
한국에 비행기로 왔어요.
2. Please speak in English.
영어로 말씀하세요.
3. John opened the door with a key.
존은 열쇠로 문을 열었다.

3. 서법조동사 변환: 영어에서 서법조동사는 'may', 'can', 'must', 'will', 'shall', 'might', 'could', 'would', 'should', 'be-able-to', 'have-to', 'be-going-to' 등을 지칭하는데, 이러한 서법조동사는 한국어로 변환될 때 어미로 표현된다. 그러나 영어의 서법조동사는 대개 여러 의미를 동시에 가지고 있어서 한국어로 변환될 때 특히 주의할 필요가 있다. 서법조동사가 가질 수 있는 의미로는 허가(Permission), 확실성/가망성 (Certainty/Probability), 가능성(Possibility), 능력(Ability), 계획(Plan), 의무/필연성 (Obligation/Necessity), 예측성(Prediction), 의지(Volition) 등이 있다.

301. 허가(Permission)
예) 1. May I borrow your pen?
당신 펜을 빌려도 좋습니까?
2. You can't play football in this park.
이 공원에서는 축구를 하지 못한다.

302. 확실성/가망성:
예) 1. He must be nearly eighty now.
그는 지금 거의 80세임에 틀림없다.
2. I think I'm going to have flu.
독감에 걸린 것 같다.
3. You'll have heard the news.
너는 아마 그 소식을 들었을 것이다.

303. 가능성:
예) 1. It may rain tomorrow.
내일은 비가 올지도 모른다.
2. It will rain before evening.
저녁 전에 비가 오겠다.
3. Can the news be true?
그 소식이 사실일까?

304. 능력:
예) 1. Can you lift this box?
이 상자를 들어올릴 수 있나?
2. I was able to help you.
너를 도울 수 있었다.

305. 계획:
예) 1. I'm going to Glasgow next week.
다음 주에 글라스고우에 간다.
2. Are you staying in London long?
런던에 오래 머물 예정인가?
3. We are to be married in May.
우리는 5월에 결혼하기로 되어 있다.

306. 의무/필연성:
예) 1. You must obey orders without question.
너는 절대로 명령에 복종해야 한다.
2. You needn't have hurried.
너는 서두를 필요가 없었다.

〈표 1〉 평가문 수준의 등급화 기준

등급	핵심 문법 사항	문장길이	어휘 수	비고
초급 (Elementary Level)	· 현재형, 과거형, 현재진행형, 근접미래구문, 단순미래구문 · 서술형, 명령형, 부정문, 의문문, 동명사구문, 단순수동구문 · 조동사구문(가능, 의무, 추측 등), 부정사구문 · 비교구문, 관계절 · 근접미래구문, 미래표현 현재구문 · 결과문, 목적문 등	15 단어 이내	1,401	중 2 영어
중급 (Intermediate Level)	· 현재완료진행구문, 과거의 습관 표현 구문, 과거완료, 과거완료진행형, 미래진행형 · 사역구문, would 사용 조건절 구문, 간접서술문, 간접의문문 · 관계부사절, 목적/이유/대조 표현 절 · 화법 수동, 진행수동 구문 · 전치사구+동명사 구문	15 ~ 20 단어	3,237	중 3 ~ 고 1 영어
고급 (Advanced Level)	· 가정법 과거구문 · 진행수동, 완료수동형 · 고급화법 구문, 양보/조건 구문 · 내포문(embedded sentence)이 2개 이상 구문 · 구나 절이 생략된 문장 · 복합 시제 구문	20 단어 이상	1,209	고 2 이상 영어

3. We had to hurry. 우리는 서둘러야 했다.
4. Do you know how to do it?
그것을 어떻게 해야 하는지 아니?

307. 예측성:

예) 1. That will be the milkman.
그 사람은 우유 배달하는 사람일 것이다.
2. They will have arrived by now.
지금쯤 그들은 도착했을 것이다.

308. 의지:

예) 1. I will write tomorrow. 내일 쓰겠습니다.
2. You shall obey my orders!
너는 내 명령을 따라야 해!

4. 명사구 변환: 구조적으로 같은 표현이라 하더라도 구성요소의 어휘 관계에 따라 한국어 변환이 달라질 수 있는 명사구를 대상으로 한다.

401. A-of-B 명사구 변환:

예) 1. The study of English is very difficult.
영어를 공부한다는 것은 매우 어렵다.
2. I don't know the country of America.
나는 미국이라고 하는 나라를 모른다.
3. It was by the legs of the table.
그것은 그 탁자의 다리 옆에 있었다.
4. She was an angel of a girl.
그녀는 천사같은 소년였다.

2.2 평가문 수준의 분류기준

한국의 영어교육, 특히 1997년부터 실시되는 초등학교 영어 교육과 중학교 영어교육은 제6차 교육과정에서 회화중심으로 바뀌었고 그 결과 순수하게 문어체에만 적용되던 번역에 직접 적용하기 어려운 면은 있으나 기본문형, 시제, 기본어휘, 문장종류 등은 등급을 정하는 기준을 제공할 수 있다. 여기에서 초등학교 영어교육은 1997년부터 시행되고 있으므로 그 전까지 영어를 처음 접하는 학생들에게 가르치던 중학교 영어교재와 중복되는 부분이 있음은 사실이나 초등학교 영어교육의 취지가 영어에 대한 동기유발이고 이를 위해 문자인식보다는 소리와 리듬인식이 주가 되는 놀이와 게임, 노래 등이 강조되므로 본격적인 문자위주의 영어교육은 여전히 중학교에서 시작된다고 볼 수 있어 기본 어휘목록표를 제외한 다른 부분은 고려 대상에 포함시키지 않았다. 결과적으로 본 연구에서는 우리나라 영어 교육의 기준[9][10][11][12]을 참조하여 만든 표 1과 같이 요약된 기준에 의해 평가문의 수준을 초급, 중급, 고급으로 분류하였다.

2.2.1 어휘의 측정

어휘의 측정은 상당부분 기계번역기에 내장된 사전의 규모에 따라 다르므로 평가의 초점을 어휘의 인지도와 함께 문맥에 맞는 어휘선택 여부, 또 올바른 목표어로 변환되었는지에 둔다. 특히 기계번역기가 인지하는 어휘의 개수보다는 개개 어휘의 정확한 이해여부, 가능성있는 대역의 산출 등이 고려되어야 한다. 예를 들어, 다의어를 어떻게 처리하는지 하는 것은 어휘의 인지도와 함께 번역기의 문맥에 대한 이해도를 측정하는데 좋은 자료가 될 수 있다.

가. 초급 어휘

영역에 상관없이 사용되는 비전문적 어휘로 일상 생활에 필요한 낱말로 구성된다. 초, 중학교 영어교과목[9][10][11]의 기본 어휘표와 시사영어사 간행 영한사전의 빈도수별 구분에 의해 가장 빈도수가 많은 *** 표시된 1,401단어를 중심으로 한다. 초등학교 기본어휘 800개와 중학교 어휘 995개 중 중복어 670여 단어, 초등학교에만 제한되어 나온 130어휘, 중학교에만 나오는 325어휘, 품사 변형에 따를 모든 어휘의 굴절어로 총 어휘 수는 1,200여개이며 이들 교과서 기본어휘와 사전의 최다 빈도수 어휘는 대부분 중복된다. 굴절형을 감안하면 초급 수준의 어휘는 1,350~1,400개이다.

나. 중급 어휘

고등학교 1학년까지의 수준에서 배우는 어휘[12]로 일상생활 영어보다 좀 더 어려운, 즉 초급어휘의 '어려운 쪽' 동의어를 포함하며, 시사영어사 영한 사전의 ** 표시된 3,237개의 어휘가 이에 해당된다. 초급어휘를 포함하면 중급 수준에서 다루어야 할 총 어휘 수는 4,700개 정도이다.

다. 고급 어휘

미국에서 대학생활을 할 수 있는 수준으로 각 분야별 기초 학문 용어를 포함하나 고도의 전문용어는 배제한다. 시사 영한사전의 *표시된 1,209개의 어휘를 중심으로 하며 초,중급의 어휘까지 포함하면 총 6,300~6,400개의 어휘가 고급 구문에 활용된다.

2.2.2 구문의 측정

구문의 측정에는 문법 사항, 시제, 문의 종류, 문장을 구성하고 있는 단어의 수에 의한 문장의 길이가 중요한 기준이 된다.

가. 초급 구문

우리나라 중학교 영어 교과서 2학년 수준까지로 한다. 주요 일상생활에 필요한, 가장 빈도수 높은 구문이 이에 속한다. 초급문은 15단어 이내로 구성된 문장으로 단문과 중문을 포함하며 문의 종류로는 평서문, 의문문, 명령문, 감탄문이 모두 포함된다. 시제상으로는 현재, 미래, 현재완료, 현재진행, 과거, 과거진행 등이 포함되며 관계대명사의 사용도 함께 고려된다. 초급문장은 주로 의미상 한가지 내용을 담고 있고, 극히 단순한 경우를 포함하고는 종속절이 포함되지 않는다.

나. 중급 구문

영어 구문을 등급화할 때 Oxford나 Cambridge 대학 출판부등 ELT 전문 출판사의 문법서[13][14][15][16]는 대개 초급과 중급으로, 혹은 초급, 초-중급, 중급으로 나누어져 사실상 영어 문법의 제 현상이 대개 중급에서 완료되고 있음을 시사해준다. 위에 언급한 초급의 기준이 좀 더 복잡하게 얽혀 있거나 시제나 구문 등이 혼합되어 나타나는 경우, 문장내 단어나 구절의 생략, 대치 등을 중급으로 설정한다. 특히 중급에 포함되어야 할 독립적 문법사항은 조건절의 등장과 완료진행, 완료진행수동구문 등이다. 우리나라 교과서 수준으로 고등학교 1학년까지의 구문이 된다. 문장의 길이는 20단어 이내가 된다.

다. 고급 구문

고등학교 2,3학년 교과서에만 등장하는 구문을 포함, 초급과 중급 문형의 시제상, 구문상, 제 현상이 다양하게 혼합되어 복잡해진 문장들을 포함한다. 예를 들면, 같은 문장 내에 여러 개의 내포문이 있는 경우, 혹은 조건절이 복잡한 시제로 표현되는 경우, 종속절과 관계절 등이 포함되어 복잡하고 어렵거나, 문장 순서의 도치, 위치 변경, 절의 생략 등을 포함한다. 문장의 길이는 20단어 이상 무제한이며 초.중급 수준을 마친 사람들만이 이해할 수 있는 구문들로 구성된다.

2.3 평가문 선정 기준과 고려사항

영어 이해능력 평가문은 기본적으로 HP-NL평가자료에 기초하고 있다. HP-NL평가문은 문법현상 중심의 핵심문장으로 구성되어 특정한 문법현상을 가장 분명히 드러낼 수 있는 어순, 내용, 낱말로 이루어진 뼈대와 같은 문장들로 이루어져 있다. 그러나, HP-NL 평가문은 잘 알려진 바와 같이 비록 문법적으로 정확하다는 구문도 실제 용례에 따른 것으로 표준문법의 기준에 맞춰 평가되지는 않은 것이여서 가끔씩 영어를 모국어로 사용하는 화자들의 감각에 비춰 보아도 어색하거나 빈도수가 극히 제한되거나, 심지어 비문법적이라고 할 수 있는 문장들이 포함되어 있다. 또 문법적으로 오류가 없다고 내놓은 912문도 다양한 텍스트에서 발췌된 살아있는 예문이 아니고 인공적으로 조립한 것이여서 실제 쓰임새와 거리가 있을 수 있다는 점과, 담화, 특히 인간과 NLP시스템 간의 상호작용까지를 포함하는 의미로서의 담화 중심의 문장구성은 기계번역기가 처리해 주기를 바라기에는 무리가 되는 면이 있다. 결론적으로 영한 번역 평가를 위한 영어 이해능력 평가문 구축을 위하여 본 연구에서는 HP-NL 평가문에 대하여 다음과 같은 확장 작업을 실시하였다.

- 인공적인 HP-NL 평가문에 대응하여 한국의 중.고교 영어 교과서의 단어 수준이나 문장 수준에 맞게 구문을 대치하거나 실제 사용 예문, 특히 실제 영어수준을 평가하는데 가장 객관적인 기준으로 인정받고 있는 TOEIC의 실제 혹은 모의문항[17][18][19]을 직접 선정하는 방법으로 HP-NL 평가문을 다양하게 확장하였다.
- 영어교육을 전공한 원어민 전문가의 언어적 직관(native intuition)에 비추어 각 문장을 '좋은 문장(good sentence)', '문법적으로 용납될 수 있는 문장(acceptable sentence)', '문맥에 따라 가능한 문장(O.K. according to the context)', '문법적으로 오류가 있는 문장(ungrammatical sentence)'으로 나누고 가능한 것은 수정하였다. 특히 HP-NL 평가문 중에서, 1) There is Abrams, Browne, and Chiang.과 같이 스쳐 지나가는 대상을 하나하나 열거할 때 쓰는 문장 혹은 특수한 상황을 전제로 하는 문장, 2) You better do it now. 처럼 지나치게 구어체적이여서 표준문법에 맞지 않는 경우, 3) Abrams showed the programmers themselves. 처럼 의미상이나 구문상으로 맞지 않는 문장, 4) Chiang hired Devito, and manages Browne. 처럼 시제가 복합되어 나오는 경우, 5) There fails to be a bookcase in the office. 처럼 어색하거나 부

적절하다고 판단된 문장 등을 모두 삭제, 또는 수정하였다. 참여한 원어민은 영어를 모국어로 하고 미국이나 영국에서 대학이상의 교육을 받은 세 명의 대학 영어회화 담당 교수이다.

- ELT전문 출판사의 문법서를 중심으로 한 표준문법에 비추어 비교, 검토함으로써 표준문법의 분류에 맞추어 재조정하고 수준별로 등급(초, 중, 고급)을 나누었다.

한편, 영한 변환능력 평가문은 우선 영어 학습자를 위한 도서[20][21][22][23][24][25][26]나 영어로 된 한국어 문법서[27]에서 예문을 수집하고, 다음과 같은 기준에 따라 선정하였다.

- 평가 대상 부분의 일반적인 한국어 번환을 명확히 살펴볼 수 있는 예문.
- 평가 대상 부분의 의미가 명시적으로 나타나면서 간결한 문장.
- 단어선택과 관련한 예문의 경우에는 의미(대역어) 구분이 명확하고 사용 빈도가 높은 다중 변환 단어들을 대상으로 품사별 비율과 유형별 의미갈래 비율을 반영하여 단어의 각 의미마다 예문을 선택함.

영어 이해능력과 영한 변환능력을 따로 측정하기가 어렵듯이 결국 영한 변환능력의 측정용 구문과 영어 이해능력의 측정용 평가문은 대부분 중복된다. 그러나 굳이 이 두 가지로 구분하는 것은 두 언어가 구조나 문법면에서 너무나 다르고 또 번역의 궁극적인 목표가 자연스러운 한국어 표현이라는 점을 감안, 특히 오류가 발생하기 쉬운, 그러면서도 두 언어간의 어법을 가장 분명히 드러낼 수 있는 일련의 제 현상을 집중적이고 반복적으로 고찰해 봄으로써 각 평가대상(시스템)의 번역 능력을 좀 더 극명하게 비교, 평가해 보고 구체적인 개선 방향을 주기 위한 것이다.

〈표 2〉 영어 이해능력 평가를 위한 번역의 정확성 평가기준

등급	번역의 정확성 평가 기준
A	모범답안의 전문가 번역문장의 의미가 기계 번역문장에 '모두(ALL)' 표현되어 있어, 기계번역문장이 모범답안의 전문가 번역문장과 의미적으로 완전히 일치한다.
B	모범답안의 전문가 번역문장의 의미가 기계 번역문장에 '대부분(MOST)' 표현되어 있어, 기계번역 문장에서 전문가 번역 문장의 의미를 쉽게 찾을 수 있으나 부적절한 어휘 선택이 한두 개 들어있다.
C	모범답안의 전문가 번역문장의 의미가 기계 번역문장에 '많이(MUCH)' 표현되어 있어, 기계번역 문장과 전문가 번역 문장이 의미적으로 상당히 가깝다.
D	모범답안의 전문가 번역문장의 의미가 기계 번역문장에 '조금(LITTLE)' 표현되어 있어, 기계번역 문장으로부터 전문가 번역 문장의 의미를 완전히 유추해 내기가 어렵다.
F	모범답안의 전문가 번역문장의 의미가 기계 번역문장에 '전혀(NONE)' 표현되어 있지 않아, 기계번역 문장으로부터 전문가 번역 문장의 뜻을 전혀 유추할 수 없거나, 번역이 되지 않은 상태로 영어 표현이 상당 부분 그대로 나타나 있다.

〈표 3〉 영어 변환능력 평가를 위한 번역의 정확성 평가기준

등급	번역의 정확성 평가 기준
O	영어 평가문의 지정 부분의 의미가 기계번역 문장에 "많이" 표현되어 있음.
P	영어 평가문의 지정 부분의 의미가 기계번역 문장에 "조금" 표현되어 있음.
X	영어 평가문의 지정 부분의 의미가 기계번역 문장에 "전혀" 표현되어 있지 않거나 영어 표현 그대로 나타나 있음.

3. 평가 기준

번역의 품질은 어떻게 평가될 수 있는가? 일반적으로 번역의 품질은 사용자와 관계없는 본질적인(intrinsic) 품질과 텍스트와 사용자에 의존적인(extrinsic) 품질이 있고, 이로 인해 번역의 품질을 평가하기 위한 절대적인 기준을 정의하기는 어렵다. 따라서 평가 기준은 번역 유형과 목적에 따라 달라질 수 있다. 예를 들어, 출판을 위한 번역의 경우에는 후편집 비용(후편집 비율과 시간)이 평가 기준이 될 수 있고, 영한 기계번역 시스템과 같이 아직 확장과 개선이 필요한 기계번역 시스템들을 평가하기 위한 기준들은 주로 언어학적이다.

언어학적 평가에는 입력(평가자료)과 출력(번역결과)만을 고려하는 블랙박스(black-box) 평가와 기계번역 시스템의 설계 과정에서 기술한 언어학적 이론에 따라 모듈별로 정확하게 처리가 되고 있는지를 검사하기 위한 유리박스(glass-box) 평가가 있는데, 본 논문에서는 실험 환경을 고려하여 블랙박스 평가 방법을 택하고 있다. 언어학적 블랙박스 평가는 번역 오류에 대한 정량적 척도(즉, 후편집자가 행한 단어 추가/삭제/대체/이동 빈도수)와 번역 품질에 대한 주관적 척도로 수행될 수 있다. 여기에서 정량적 척도에 의한 평가는 초벌 번역의 정확성이나 문체의 적절성에 대한 판단이 후편집자나 분야에 따라 다를 수 있기 때문에 대규모 실험 없이는 평가의 객관성을 확보하기가 어렵다[5]. 이로 인해 번역의 언어학적 평가에는 정확성(accuracy)과 이해도(intelligibility) 측정과 같이 사람의 직관적 판단에 의존하는 주관적 척도로 수행되는 평가가 아직까지 유효하며[1][3][5][6][29], 본 논문도 이러한 점을 감안하여 이것을 평가에 사용하였다.

3.1 번역의 정확성

번역에 있어서 정확성이란 "번역된 텍스트가 원 텍스트와 같은 정보를 담고 있는 정도"[5][28]를 함축적으로 나타낸 말이다. 이것을 측정하기 위한 방법으로는 번역 텍스트를 다시 원문으로 번역시킨 다음 원문과의 불일치 비율을 측정하여 평가하는 역번역(back-translation) 측정 방법, 번역 매뉴얼로 작업을 수행하였을 때 생기는 작업 실수를 측정하여 평가하는 작업성취도 측정 방법, 평가자의 직관적 판단에 의한 등급 부여 방법 등이 있다. 이 중에서 역번역 측정 방법은 양방향 번역이 가능한 시스템에서만 가능하고, 작업성취도 측정 방법의 경우에는 시현하여 평가하기가 어려운 문제를 가지고 있다. 그래서 본 논문에서와 같이 대개의 경우 다수 평가자에게 등급을 부여하게 하여 통계적 분석을 수행하는 설문조사 방법을 취하고 있다.

본 연구에서 정확성 평가는 전문가가 만든 평가문(영어)의 모범번역문(한국어)에 표현된 의미가 평가대상 번역문(한국어)에 어느 정도 나타나느냐를 측정하는 것으로 하였다. 또 등급 설정은 다음 두 가지 점을 고려하였다. 즉, 영한 기계번역의 영어 이해능력 평가에서는 각 평가문에 대하여 기계번역 시스템이 자유롭

〈표 4〉 번역의 이해도 평가 기준

등급	번역의 이해도 평가 기준
A	읽기가 '아주 좋음(EXCELLENT)': 문법 및 문체의 부적절함이 전혀 없어 번역 문장을 보통의 한국어 텍스트 문장처럼 읽을 수 있다.
B	읽기가 '좋음(GOOD)': 사소한 문법적, 문체적 부적절함(예, 어색한 단어선택이나 구문배열)은 있으나 전체적으로 보아 번역 문장이 명확하고 이해하기 쉽다.
C	읽기가 '대체로 괜찮음(FAIR)': 부적절한 단어선택이나 구문배열, 번역되지 않은 단어 등이 나타나 번역 문장을 여러 번 반복해서 읽어야만 이해가 가능하다.
D	읽기가 '어려움(POOR)': 주요 단어가 번역되지 않은 채 남아 있어 일반적인 상식을 가지고는 번역 문장을 이해할 수 없고, 수없이 반복해서 읽어야만 겨우 의미를 추측할 수 있다.
F	읽기가 '아주 어려움(VERY POOR)': 전혀 번역이 되지 않아 원문 그대로거나, 번역된 문장을 여러 번 읽어도 무슨 의미인지 전혀 이해할 수 없다.

게 답(번역)할 수 있고 그 답을 모범답안(전문가 번역문)과 비교하여 전체적으로 평가한다는 점에서 일반적인 논술형 평가와 같이 5등급으로 평가하는 것이 타당할 것이라는 점과, 영한 번환능력 평가는 영어 평가문의 지정된 부분이 얼마나 정확하게 한국어로 변환되었는가를 평가하기 때문에 사지선다형 평가와 논술형 평가의 중간 정도라 할 수 있어 3등급으로 평가하는 것이 좋겠다는 점을 고려하여 등급을 설정하였다. 한편, 평가 측정의 객관성을 확보하기 위하여 같은 평가항목을 3명의 평가요원이 평가하게 하여 그들이 부여한 등급의 평균값을 그 평가항목의 평가 결과로 사용하는 최소 형식을 갖는 설문조사 방법을 사용하였다. 결과적으로 본 연구에서 번역의 정확성은 평가자가 평가대상 번역문과 전문가 번역문의 비교 결과에 대하여 부여한 등급으로 나타날 것이다.

〈표 2〉의 등급 기준은 DARPA의 5등급 평가 기준[29]에 기초하여 만든 것으로, 영어 이해능력 평가를 목적으로 번역의 정확성에 대한 평가자의 주관적 판단을 등급화하기 위한 것이다. 이 평가 기준은 평가요원이 각 영어 이해능력 평가문에 대해 전문가 번역문의 의미가 평가 대상자(시스템)의 번역문에 얼마나 나타나느냐를 판단하여 'A:모두(ALL)', 'B:대부분(MOST)', 'C:많이(MUCH)', 'D:조금(LITTLE)', 'F:전혀(NONE)' 등급 중 하나를 적절히 선택할 수 있도록 돕기 위한 것이다.

〈표 3〉의 평가 기준은 영한 변환능력을 평가하기 위한 것이다. 영한 변환능력의 평가 결과는 각 평가문의 특정 부분이 얼마나 정확하게 한국어로 변환되었느냐에 따라 '만족(O)', '부족(P)', '실패(X)' 등급으로 나타날 것이다.

3.2.2 번역의 이해도

번역의 이해도는 "독자가 번역 문장을 얼마나 쉽게 이해할 수 있는가 혹은 번역 문장이 얼마나 명확한가의 정도"[5]를 나타내는 것으로, 문법 오류, 오역, 번역되지 않은 단어 등에 의해 달라질 수 있다. 따라서 번역의 이해도는 정확성 평가와 달리 비교할 수 있는 전문가의 번역문이 없이도 한국어를 모국어로 쓰는 평가자의 한국어 사용법(문법, 철자법, 어휘선택 등)에 전적으로 의존하여 평가할 수 있다. 이것을 평가하기 위한 방법으로는 정확성 평가에서 사용한 등급 부여 방법과 번역 텍스트에서 다양한 단어들을 감춘 다음 적절하게 채우도록 요청하여 평가하는 Close 테스트 방법이 있는데, 본 논문에서는 작업 편의성과 효율성 때문에 등급 부여 방법을 택하여 사용하였다. 결과적으로 본 연구에서 번역의 이해도는 표 4의 기준에 의해 평가되었다. 〈표 4〉는 번역의 이해도에 대한 평가자의 주관적 판단을 등급화하기 위한 판단 기준으로, DARPA의 유창성 평가 기준을 참조하여 만들었다.

4. 영한 번역 평가 실험과 분석

4.1 평가 실험 준비

4.1.1 평가 대상문 선정

본 실험의 목적은 제안한 평가방법이 실제 영한 번역의 언어학적 평가에 효과적인가를 분석하는데 있기 때문에, 평가 대상문은 한 시스템의 언어학적 적용범위를 다양하게 측정해 볼 수 있도록 선정해야 한다. 또 HP-NL 평가자료를 기초로 구축한 평가자료가 잘 구축되었는가를 가늠해 보는 것도 이 실험에서 살펴보아야 한다. 이를 고려하여 실제로 실험에 사용할 평가문은 평가자료로부터 아래와 같이 구성하였다.

- 영어 이해능력 측정을 위한 HP-NL 평가문: 평가자료의 HP문 중에서 영어 사용 원어민이 일반적인 영어 문장이라고 검토한 771문장(초급 427문, 중급 280문, 고급 64문)을 대상으로 함.
- 영어 이해능력 측정을 위한 HP-NL' 평가문: 인공적인 HP문에 대응하여 실제 사용 예문을 수집하여 구성한 평가자료 중에서 영어 사용 원어민이 일반적인 영어 문장이라고 검토한 735문장(초급 393문, 중급 281문, 고급 61문)을 대상으로 함.
- 영한 변환능력 측정을 위한 평가문: 시범적으로 구축한 영한 변환능력 평가문 전체(299문장)를 대상으로 함.

결과적으로 본 실험을 위해서 평가문은 3종류 총 1,805문장이 평가자료로부터 선정되었다.

4.1.2 시스템 번역 시험

시스템 번역 시험은 앙꼬르2.0, 워드체인지3.0, 트래니96(전문가용) 등 3종의 영한 기계번역 시스템을 대상으로 앞 절에서 선정한 평가 대상문에 대하여 동일한 컴퓨터 환경에서 시행하였다. 각 시스템의 출력결과는 일련번호와 함께 인쇄되어, 나중에 평가요원에 의해 각 문장별로 평가가 이루어질 수 있도록 하였다.

4.1.3 평가요원 선발

평가요원은 앞에서 제시한 평가기준에 따라 영한 기계번역 번역문에 대하여 등급을 적절히 부여할 수 있는 한국어를 모국어로 하는 대학수학능력을 가진 사람들을 표 5와 같이 선발하였다.

〈표 5〉 평가요원 선발

구 분		정확성 평가	이해도 평가	평가요원의 수
영어 이해능력 평가문	HP-NL문	3명	3명	9명
	HP-NL'문	3명		
영한 번환능력 평가문		2명		2명

평가요원은 평가문과 평가 유형에 따라 달리 선발하였다. 영어 이해능력 평가는 영어 문장 전체 의미가 제대로 한국어로 얼마나 정확하고 이해할 수 있게 표현되었는가를 측정하는 것이고, 반면에 영한 변환능력 평가는 영어 문장의 특정 부분이 얼마나 제대로 한국어로 정확하게 표현되었는가를 판정하는 것이다. 따라서 영한 변환능력 평가문의 경우에는 이해도를 평가하지 않기 때문에 평가요원을 선발하지 않았다. 또, 정확성 평가도 2명의 평가요원이 각각 반씩 나누어 수행하였다. 그리고, 영어 이해능력 평가문의 경우에는 정확성 평가가 이해도 평가보다 힘들다고 판단하여 정확성 평가요원을 이해도 평가요원보다 더 많이 선발하여 작업 부담의 균형을 맞추려고 노력하였다.

4.2 평가 실험 결과

4.2.1 영어 이해능력 평가 결과

영어 이해능력 평가문에 대하여 실험한 결과, 실험 대상 영한 기계번역 시스템들은 번역의 정확성과 이해도 측면에서 〈표 6〉, 〈표 7〉과 같은 정도의 번역 능력을 보이는 것으로 나타났다. 〈표 6〉과 〈표 7〉에서 각 시스템의 평가 결과는 문장별로 3명의 평가요원이 부여한 등급을 평균한 값을 나타낸다.

〈표 6〉과 〈표 7〉의 결과를 보고 알 수 있듯이, 실험 대상 영한 기계번역 시스템들은 번역의 정확성이 HP-NL문 3.2~3.7 등급과 HP-NL'문 3.0~3.7등급이고, 번역의 이해도는 HP-NL문 2.1~2.7 등급과 HP-NL'문 2.1~2.8등급을 보이고 있다. 이러한 결과를 단순히 백분율로 나타내 보면, 번역의 정확성은 HP-NL문 46~56%{(6-3.7)*100/5~(6-3.2)*100/5}와 HP-NL'문 46~60%{(6-3.7)*100/5~(6-3.0)*100/5}이고, 이해도는 HP-NL문 66~78% {(6-2.7)*100/5~(6-2.1)*100/5}와 HP-NL'문 64~78%{(6-2.8)*100/5~(6-2.1)*100/5}이다. 실험 결과만을 볼 때 앙꼬르가 다른 시스템에 비해 우위를 보이고 있다. 그러나, 그 차이가 크지 않고, 또 번역의 능력도 기대보다 낮기 때문에, 시스템 간의 순위는 별 의미가 없다고 할 수 있다.

또한 초급수준만을 놓고 볼 때, 실험 대상 시스템들은 HP-NL문 56~60%(3.0~3.2등급)와 HP-NL'문 56~68%(2.6~3.2등급)의 정확성이 있어, 중급수준 HP-NL문 36~50%(3.5~4.2등급)와 HP-NL'문 36~54%(3.3~4.2등급)이나, 고급수준 HP-NL문 28~40%(4.0~4.6등급)와 HP-NL'문 32~40%(4.0~4.4등급)에 비해 높음을 알 수 있다. 번역의 이해도도 초급문장에 대해서는 HP-NL문과 HP-NL'문이 모두

〈표 6〉 HP-NL 평가문에 의한 평가 결과

	평가 문장	번역의 정확성			번역의 이해도		
		워드체인지	트래니	앙꼬르	워드체인지	트래니	앙꼬르
초급	427문	3.0등급	3.2등급	3.0등급	2.1등급	2.5등급	1.9등급
중급	280문	3.7등급	4.2등급	3.5등급	2.5등급	3.0등급	2.2등급
고급	64문	4.4등급	4.6등급	4.0등급	2.8등급	3.2등급	3.2등급
합계	771문	3.4등급	3.7등급	3.2등급	2.3등급	2.7등급	2.1등급

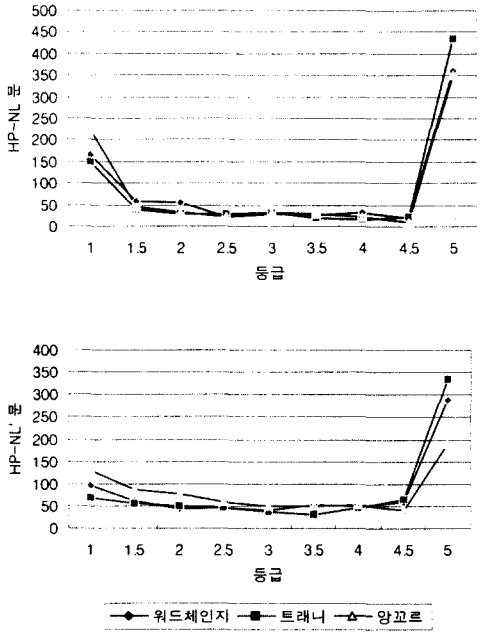
〈표 7〉 HP-NL′ 평가문에 의한 평가 결과

	평가 문장	번역의 정확성			번역의 이해도		
		워드체인지	트래니	앙꼬르	워드체인지	트래니	앙꼬르
초급	393문	3.1등급	3.2등급	2.6등급	2.2등급	2.5등급	1.9등급
중급	281문	3.8등급	4.2등급	3.3등급	2.6등급	3.1등급	2.3등급
고급	61문	4.4등급	4.4등급	4.0등급	2.9등급	3.4등급	2.9등급
합계	735문	3.5등급	3.7등급	3.0등급	2.4등급	2.8등급	2.1등급

70~82%(1.9~2.5등급)을 보이고 있어, 중급수준 HP-NL문 60~76%(2.2~3.0등급)와 HP-NL′문 58~74%(2.3~3.1등급)이나, 고급수준 HP-NL문 56~64%(2.8~3.2등급)와 HP-NL′문 52~62%(2.9~3.4등급)에 비해 좋은 결과가 나옴을 알 수 있다. 이러한 평가 결과는 평가자료의 수준별 분류가 어느 정도 타당성 있게 구성되었음을 내보이는 결과라 할 수 있다.

HP-NL′문 평가 결과가 HP-NL문 평가 결과보다 정확성 평가에서 4~8%, 이해도 평가에서 0~4% 더 좋게 나타났다. 이러한 차이는 무엇보다도 인공적인 HP-NL문이 실제적인 HP-NL′문에 비해 비록 사용된 단어 수가 적긴 하지만 특정 단어들만을 사용하여 다양한 문법 현상을 표현하고 있기 때문에, 사전에 단어의 모든 의미와 쓰임을 갖고 있지 않는 번역 시스템들에게는 인공적인 HP-NL문이 실제적인 HP-NL′문보다 오히려 번역 결과가 나쁘게 나타날 수 있었다. 따라서 HP-NL′문과 그 모체가 되는 HP-NL문은 서로 보완될 필요가 있으며, 본 실험에서와 같이 둘 다 평가에 이용하여 평가의 신뢰구간을 확보하는 것이 필요하다 할 수 있다.

앞에서 살펴본 바와 같이, 실험 대상 영한 기계번역 시스템은 일반 사용자가 사용하기에는 정확성이 많이 부족한 것처럼 보인다. 즉, 실험 대상 시스템 모두 번역의 정확성을 높이기 위한 개선이 무엇보다 요구된다. 사실 인터넷 정보화 시대에서 기계번역은 이해도보다 정확성이 더 강조되고 있는 것이 사실이고, 현재 기술로는 양질의 완전 자동번역이 기대하기 힘들 뿐만 아니라, 정보 습득을 위한 인터넷 정보의 기계번역은 원 내용이 사용자에게 얼마나 정확히 전달되었느냐가 더 중요한 평가요소로 인식되고 있다. 이와 관련한 실험 대상 시스템의 개선점을 보기 위하여, 실험 결과에 대하여 평가요원이 어떠한 양상으로 정확성을 평가했는가를 분석해 보도록 하자. 이러한 분석은 기본적으로 평가요원이 정확한 번역이라고 평가한 문장의 수와 완전히 실패한 번역이라고 평가한 문장의 수에 따라 대상 시스템의 개선 정도를 실측해 볼 수 있다는 전제에 기반하고 있으며, 그 조사 결과는 (그림 1)에 요약되어 있다.

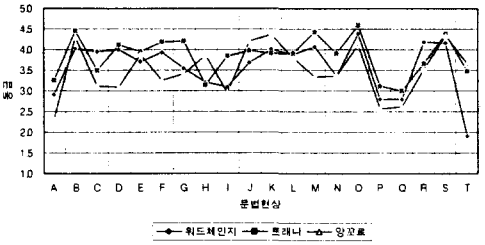


(그림 1) 평가문에 대한 평가요원의 평가 결과

(그림 1)은 실험 대상 영한 번역 시스템들이 출력한 HP-NL문과 HP-NL′문의 번역 결과에 대하여 평가요원이 각각 번역의 정확성을 문장별로 측정한 결과를 종합적으로 나타낸 것이다. 그림에 나타나 있듯이, HP-NL과 HP-NL′문에 대하여 실험 대상 영한 번역

시스템들은 정확하게 번역한 문장 수보다 완전히 틀리게 번역한 문장의 수가 더 많은 것으로 나타나 많은 개선이 요구되는 것으로 조사되었다. HP-NL문의 경우 앙꼬르는 정확하게 번역한 문장의 수가 225문이나 되어 워드체인지 165문, 트래니 148문보다 더 많았지만, 완전히 틀리게 번역한 문장의 수는 이보다 훨씬 더 많은 것으로 나타나 앙꼬르가 351문, 워드체인지 361문, 트래니 435문으로 조사되었다. 즉, 앙꼬르가 잠재적으로 수정해야 할 문장의 수는 1등급 평가를 받은 225문을 제외한 546문(70.8%)이고, 완전히 틀리게 번역하여 시급히 확장할 필요성이 있는 문장의 수는 5등급 평가를 받은 351문(45.5%)인 것으로 나타나, 워드체인지의 잠재 수정 문장 606문(78.6%)과 확장 요구 문장 361문(46.8%)이나, 트래니의 잠재 수정 문장 623문(80.8%)과 확장 요구 문장 435문(56.4%)에 비해 상대적으로 우수한 것으로 나타났다. 한편, HP-NL'문의 경우에도 결과는 비슷하게 나타났다.

번역 시스템의 개선이나 확장 방향은 평가 결과를 문법현상별로 살펴보면 알 수 있다. (그림 2)는 위의 실험 결과를 전체적으로 살펴보기 위하여 대분류 문법현상에 따라 그래프로 표현하였다.



(그림 2) 대분류 문법현상별 평가 결과 분석 그래프

(그림 2)에서는 간결성을 위해 문법현상별로 HP-NL문과 HP-NL'문의 평가 결과를 평균하여 나타내었다. 시스템의 개선 방향을 개략적으로 살펴보기 위하여 먼저 3.5등급을 '개선 임계값'이라고 하였다. 즉, 3.5등급보다 좋지 못한 평가를 받은 문법현상들은 시스템이 당장 개선해야 할 대상이고, 3.5등급보다 좋은 평가를 받은 문법현상들은 당장 개선을 하지 않아도 된다고 보았다. 이를 토대로 실험 대상 영한 번역 시스템의 특징을 살펴보면 다음과 같다.

- 모든 시스템이 동시에 개선 임계값보다 좋게 평가받은 문법현상은 어휘 의존도(A.), 부정문(P.), 시제와 상(Q.) 등 단지 3개 항목(15%)뿐임.

〈표 8〉 영한 변환능력 평가문에 의한 평가 결과

문법현상	문장	평가	워드체인지	트래니	앙꼬르
1. 어휘 변환	148문	O(만족)	56	45	66
		P(부족)	7	10	7
		X(실패)	85	93	75
		정확률	40%	34%	47%
2. 격 변환	65문	O(만족)	51	49	50
		P(부족)	9	6	9
		X(실패)	5	10	6
		정확률	85%	80%	84%
3. 서법조동사 변환	77문	O(만족)	42	36	56
		P(부족)	19	20	12
		X(실패)	16	21	9
		정확률	67%	60%	81%
4. 명사구 변환	9문	O(만족)	2	1	2
		P(부족)	5	5	3
		X(실패)	2	3	4
		정확률	50%	39%	39%
합계	299문	O(만족)	151	131	174
		P(부족)	40	41	31
		X(실패)	108	127	94
		정확률	57%	51%	63%

- 모든 시스템이 동시에 개선 임계값보다 좋지 못한 평가를 받은 문법현상은 평서문 위치 이동(B.), 비교상관구문(E.), 외치구문(J.), 명령구문(K.), 목적어 공백 보문(L.), 조건문(O.), 서술 부가어(R.), 삽입문(S.) 등 8개 항목(40%)이나 됨.
- 개선 임계값보다 좋게 평가받은 문법현상의 수를 시스템별로 살펴보면, 워드체인지가 6개(30%), 트래니 5개(25%), 앙꼬르 10개(50%)로 나타나, 상대적으로 트래니와 워드체인지가 앙꼬르보다 더 많은 개선이 요구되고 있음.

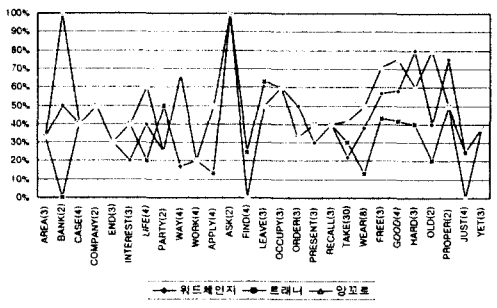
4.2.2 영한 변환능력 평가 결과

실험 대상 영한 기계번역 시스템이 보인 영한 변환 능력은 〈표 8〉과 같다. 표 8에서 영한 변환능력의 평가는 각 평가문의 지정된 부분이 얼마나 정확하게 번역되었느냐에 따라 O(만족), P(부족), X(실패)로 판정하여 영한 변환의 대분류 평가항목과 시스템별로 집계한 결과이다. 여기에서 정확률은 각 시스템이 'O(만족)'의 판정을 받은 문장들을100% 정확하게 변환한 것이라 하고, 'P(부족)'의 판정을 받은 문장들에 대해서 50%만이 정확하게 변환한 것이고 'X(실패)' 문장들에 대해서는 변환이 전혀 정확하지 않다고 가정할 경우의 값이다.

〈표 8〉의 결과를 보면, 현재 영한 기계번역 시스템의 영한 변환 능력은 51~63% 수준으로, 영어 이해능력의 초급과 비슷한 2.9~3.5등급 수준임을 알 수 있다. 또 영어 이해능력 평가와 함께 영한 변환능력도 앙꼬르가 가장 우위를 보인 것으로 나타났다. 한편, 평가 항목별로 나타난 결과를 보면 모든 시스템들이 '격 변환(83%)'과 '서법조동사 변환(69%)'은 대체로 잘 번역하였지만, '어휘 변환(40%)'과 '명사구 변환(43%)'은 상당히 저조한 정확률을 보여 이에 대한 개선이 많이 요구되는 것을 알 수 있다. 결과적으로, 앞에서 언급한 시스템 개선 임계값 3.5등급(50%)보다 좋지 못한 정확률을 보이고 있는 문법현상, 특히 어휘변환의 개선에 관심을 가질 필요가 있다.

어휘변환의 개선점을 살펴보기 위하여 실험 결과를 (그림 3)에 요약하였다. (그림 3)에서 알 수 있는 사실은 다음과 같다.

- 어휘 변환의 평균 정확률은 34~47%에 불과하지만, 동사 'ASK'에 대해서는 모든 시스템이 정확률 100%를 보이고 있음. (참고로 동사 'ASK'는 '묻다'와 '요청하다'로 두갈래 변환이 가능하기 때문에 둘 중 어느 하나만을 무조건 선택할 경우 확률은 50%임.)
- 반면에 동사 'FIND'에 대해서는 25%의 정확률을 보이고 있는 트래니를 제외하고는 워드체인지와 앙꼬르가 전혀 정확하게 변환하지 못하는 것으로 나타남. (참조로 동사 'FIND'는 '찾다', '알다', '판결을 내리다', '제공하다' 등 네갈래 변환이 가능하기 때문에 단순 확률은 25%임.)
- '은행'과 '둑'으로 두갈래 변환이 가능한 명사 'BANK'에 대해서는 시스템별로 차이가 나타나, 워드체인지 50%, 트래니 0%, 앙꼬르 100%의 정확률을 보이고 있음.
- '방법', '길', '습관', '점' 등 네갈래 변환이 가능한 명사 'WAY'에 대해서는 워드체인지와 앙꼬르가 67%의 높은 정확률을 보이고 있지만, 트래니는 정확률이 17%에 불과함.
- 단순 확률이 5%도 채 안될 정도로 아주 다양한 변환이 가능한 동사 'TAKE'에 대해서는 워드체인지 22%, 트래니 30%, 앙꼬르 42%로 비교적 높은 정확률을 보이고 있음.
- 위와 같은 사실로 미루어 보아, 실험 대상 영한 기계번역 시스템들은 어휘 변환을 무조건적인 확률 선택이 아닌 알고리즘에 의해 처리하는 것으로 보이나, 나타난 결과로 보면 어휘별로 이에 대한 개선이 상당히 필요한 것으로 보임.



(그림 3) 어휘 변환 평가 결과 분석 그래프

5. 결론

본 논문에서는 영한 번역의 언어학적 평가요소를 반영하여 만든 평가자료에 기반하여 영한 기계번역의 질적 평가를 수행하는 모델에 관하여 기술하였다. 본 연구의 의의는 다음과 같다.

- **학제간 연구 결과.** 영한 기계번역의 품질을 측정하기 위한 평가문은 외국어로 영어를 교육하고 있는 우리나라의 환경에서 영어에 대한 언어학적인 제 고찰을 바탕으로 작성될 수 있고, 또 그런 토대 위에서 이루어진 평가문은 기존의 코퍼스만을 이용한 평가문이나 인위적으로 작성된 평가문보다 교육현장의 산물로서 교육현장에 더 가까이 갈 수 있다는 점에서 실용적이고 쉽게 적용할 수 있는 장점이 있다. 이러한 장점은 학제간 연구가 아직은 생소한 우리나라의 교육풍토에서의 어려움을 극복하고도 남을 것으로 판단되며, 특히 이를 통해 서로 다른 학문 분야에 종사하는 전공자들이 동일 목표를 위한 공동 노력의 장을 형성하는 것은 앞으로의 학문 발전에 큰 모티브를 제공할 수 있을 것이다. 특히 인터넷을 통한 국제적 의사소통이 활성화되는 이 시점에 한영 혹은 영한 번역의 수요가 급격히 대두되고 있는 만큼 영어학과 국어학, 전산학의 학제간 상호협력과 공동연구는 더욱 절실히 요구된다고 할 수 있다. 또 그런 의미에서 이번의 연구는 충분히 의미있는 시작이 될 수 있을 것이다.
- **영한 기계번역 품질의 고급화 촉진 기반 제공.** 일반적으로 평가는 현재의 전반적 능력을 측정하는 것이다. 또한 평가는 능력 향상에 필요한 개선 방향을 제시할 수 있어야 한다. 이러한 평가에서 평가 문항은 능력 향상을 촉진시킬 수 있는 기반이다. 이와 같이 영한 기계번역 평가에서 평가자료는 영한 기계번역기의 전반적 능력을 측정하고 번역품질을 개선시킬 수 있는 방향을 제공할 수 있다. 4장에서 살펴본 영한 기계번역의 평가 실험에서 알 수 있듯이 실험 대상 영한 기계번역기들은 인터넷을 항해하면서 접하는 다양한 영어 정보를 제대로 번역하기에는 아직 능력이 많이 부족한 상태이다. 이러한 상태에서 서로 비교하여 어느 영한 기계번역기가 더 우수한가를 단순히 가리는 평가는 별의미가 없다. 대신에 영한 기계번역기의 전반적 능력을 측정하고 개선점들을 점검하여 영한 기계번역 품질의 고급화를 지속적으로 촉진할 수 있는 평가가 중요하다 하겠다. 그런 의미에서 본 연구는 이러한 평가를 시행할 수 있는 기틀을 제공하고 있는 점에서 의의가 있다고 할 수 있을 것이다.

앞으로 남은 연구 과제는 우선 평가자료의 활용을 극대화하고 평가자료를 꾸준히 보완, 발전시켜 관련 분야에 종사하는 다수가 인정할 수 있는 객관성과 신뢰성을 확보하는데 있다고 할 수 있다. 이를 위해서는 무엇보다도 정기적으로 영한 기계번역에 대한 평가를 시행하고 분석하는 일이 진행되어야 할 것이다. 우선은 1998년 이후에 나온 영한 기계번역기들(인가이드, Etran2000, ClickQ, EZReader, 미래 번역기 등)을 대상으로 평가를 시행하여 1997년까지의 영한 기계번역 평가 결과와 비교하여 2~3년 사이에 어떤 변화가 있었는지 분석할 필요가 있다. 이를 통하여 좀 더 효과적이고 실제적으로 고품질 영한 기계번역기 개발에 도움을 줄 수 있고, 평가자료의 개선점들을 효과적으로 수집하고 보완할 수 있을 것이다. 또한, [30]에서와 같이 최근에 수행한 다른 연구 결과와의 비교를 통해 평가자료를 보완하는 것도 앞으로 남은 과제일 것이다.

참고 문헌

[1] Choi, Key-Sun, Seungmi Lee, Hiongun Kim, Deok-Bong Kim, Cheoljung Kweon, and Gilchang Kim(1994), An English-to-Korean Machine Translator: MATES/EK, COLING-94, 129-133.

[2] 김태주(1997), 기계번역 각 프로그램 비교 분석(영한 번역기 - 앙꼬르, 워드체인지, 트래니), 번역의 세계, 1-2:8, 18-59.

[3] 이민행, 지광신, 정소우(1998), 기계번역 시스템 측정 장치 연구, 언어와 정보, 185-219.

[4] Arnold, D, R. L. Humphreys, and L. Sadler(1993), Evaluation: An Assessment, Machine Translation, 8:1-2, 1-24.

[5] Jones, K. S. and J. R. Galliers(1995), Evaluating Natural Language Processing Systems: An Analysis and Review, Springer.

[6] 시정곤, 김원경, 고창수(2000), 영/한 기계번역 성능 평가 방안 연구, 학술진흥재단 연구결과보고서.

[7] Arnold, D, D. Moffat, L. Sadler, and A. Way(1993), Automatic Test Suite Generation, Machine Translation, 8:1-2, 29-38.

[8] Flickinger, D., M. Friedman, M. Gawron, J. Nerbonne, C. Pollard, G. Pullman, I. Sag, and T. Wasow(1987), Toward Evaluation of NLP Systems: HP-NL Test Suite, Special

Session at the 25th Annual Meeting of the Association for Computational Linguistics, 1-31.
[9] 교육부(1996), 초등학교 교육과정 해설(IV): 영어.
[10] 교육부(1993), 중학교 교육과정.
[11] 교육부(1994), 중학교 영어과 교육과정 해설.
[12] 교육부(1995), 고등학교 외국어과 교육과정 해설(I).
[13] Coe, Norman(1995), Grammar Spectrum 3: English Rules and Practice Intermediate, Oxford Univ. Press.
[14] Harrison, Mark(1995), Grammar Spectrum 2: English Rules and Practice Pre-intermediate, Oxford Univ. Press.
[15] Murphy, R.(1990), Essential Grammar in Use, Cambridge Univ. Press.
[16] Peterson, Ken(1995), Grammar Spectrum 1: English Rules and Practice Elementary, Oxford Univ. Press.
[17] 서울대 어학연구소(1996), 영어문제집, 탐구당.
[18] 시사영어사(1995), 시사 Practice Book for the TOEIC: 실전문제, YBM 시사영어사.
[19] Lin Lougheed(1989), Longman Preparation Series for the TOEIC Test: Introductory Course, Longman.
[20] 박성순(1988), BELT 영어 연구, 양영각.
[21] 동아출판사(1997), 수능형 제넥스 영한 사전, 동아출판사.
[22] 교연출판사(1996), 영단어 탐구여행, 교연출판사.
[23] 이기동(1995), 영어 동사의 의미 상/하, 한국문화사.
[24] 이주영(1996), 동사를 알면 영어가 된다 1, 20세기 플러스.
[25] 이찬승(1996), 능률 독해 Vocabulary, 능률영어사.
[26] 정재형(1996), 고교단어 강의노트, 디딤돌.
[27] Ihm, Ho-Bin, Kyung-Pyo Hong, and Suk-In Chang(1988), Korean Grammar for International Learners, Yonsei Univ. Press.
[28] Lehrberger, John and L. Bourbeau(1988), Machine Translation, Vol. 15: Linguistic Characteristics of MT systems & General Methods of Evaluation, John Benjamins Pub. Co.
[29] O'Connell, T. A. and J. White(1996), Machine Translation Evaluation, AMTA TUTORIAL.
[30] KORTERM(2001), 기계번역시스템 평가 모델링 방법과 시범 DB 구축 방법론 및 제품 분석 보고서.